Published online 30 August 2010

Nucleic Acids Research, 2011, Vol. 39, No. 1 ***19–29***
doi:10.1093/nar/gkq768

Efficient use of accessibility in microRNA target prediction

Ray M. Marín and Jiří Vaníček*

Laboratory of Theoretical Physical Chemistry, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Received April 28, 2010; Revised August 10, 2010; Accepted August 12, 2010

ABSTRACT

Considering accessibility of the 3′UTR is believed to increase the precision of microRNA target predictions. We show that, contrary to common belief, ranking by the hybridization energy or by the sum of the opening and hybridization energies, used in currently available algorithms, is not an efficient way to rank predictions. Instead, we describe an algorithm which also considers only the accessible binding sites but which ranks predictions according to over-representation. When compared with experimentally validated and refuted targets in the fruit fly and human, our algorithm shows a remarkable improvement in precision while significantly reducing the computational cost in comparison with other free energy based methods. In the human genome, our algorithm has at least twice higher precision than other methods with their default parameters. In the fruit fly, we find five times more validated targets among the top 500 predictions than other methods with their default parameters. Furthermore, using a common statistical framework we demonstrate explicitly the advantages of using the canonical ensemble instead of using the minimum free energy structure alone. We also find that 'naïve' global folding sometimes outperforms the local folding approach.

INTRODUCTION

MicroRNAs (miRNAs) are small single-stranded RNAs of ~22 nucleotides (nt) that bind to the 3′-untranslated region (3′UTR) of mRNA transcripts, usually downregulating the expression of the corresponding protein (1). Due to the important role that miRNAs play in cell differentiation, development, cancer and other biological processes in species ranging from viruses to humans (1), the identification of miRNA targets is essential. Although indirect experimental approaches to identify targets in large genomes exist (2–4), they are highly demanding in terms of resources and time. Accurate computational methods for predicting functional miRNA-3′UTR pairs are therefore necessary to guide experiments. In plants, an almost perfect base pair complementarity of the whole miRNA to the messenger is required, making target identification a simple task for standard bioinformatic tools (5). In animals, on the other hand, only a partial complementarity is necessary (6), making target prediction non-trivial (7).

Different approaches to model the miRNA-3′UTR recognition have led to several miRNA target prediction methods in animals (8–10). Existing methods have motivated successful experiments, but their reliability is still far from perfect. Most methods require a full complementarity to the so-called 'seed' region (at least six consecutive nts in positions 2–7 in the miRNA). Various authors have attempted to increase the precision of target predictions in several ways.

First, the free energy of pairing between the miRNA and the 3′UTR (so-called duplex or 'hybridization energy') was implemented to select the miRNA-3′UTR pairs with the strongest interaction, i.e. only sites with the hybridization energy below a certain cutoff (11–17). Some algorithms use such energies to rank the final set of predictions. However, this procedure assumes implicitly that the stronger the physical interaction, the more likely it is that such a pair will be functional. Several arguments have been raised against this assumption (6,18,19). In this work we show clearly that once a perfect complementarity to the seed is required (positions 2–8), ranking according to hybridization energy does not gain much. We find that such a ranking performs similarly to random ordering.

Another filter, the conservation among 3′UTRs, was implemented in most methods (12,13,15,17,20), due to the observation that functional binding sites tend to be in conserved regions of the 3′UTRs (21,22). Although conservation is probably the most powerful criterion for target prediction, increasing the precision in this way may sometimes be inconvenient because certain miRNAs have evolved to act on non-conserved 3′UTRs (18,21).

*To whom correspondence should be addressed. Tel: +41 21 693 4736; Fax: +41 21 693 9755; Email: jiri.vanicek@epfl.ch

The third approach to increase the precision of predictions is considering the accessibility of the binding sites (18,23–29). To facilitate interaction, both the miRNA and the 3′UTR should be accessible, at least in the region corresponding to the seed. However, it is only the accessibility of the 3′UTR that needs to be assessed because, in their active state, miRNAs are assembled into the RNA-induced silencing complex (RISC) that guarantees the accessibility of the miRNA seed (30). Two important factors must be considered when using accessibility: first, a criterion for selecting accessible binding sites; and second, the way final predictions are ranked. To the best of our knowledge, two general strategies have been used: (i) Some methods select only those binding sites that are 'partially accessible', i.e. sites containing stretches of three or four unpaired nucleotides, thereby enabling a more complete pairing along the whole miRNA. The predicted targets are then ranked according to an *ad hoc* score (18) or free energies (23). (ii) Other methods consider the sum of two energetic contributions: the free energy required to make the complementary site accessible (the so-called 'opening energy') and the hybridization energy (24–26). The sum (which we will refer to as the 'total free energy') is used not only to decide which binding sites are accessible but also to rank the predictions. However, as the currently available folding algorithms cannot account for RNA–protein interactions, ranking by total free energy is expected to fail if such interactions make a significant contribution to the final energy balance. In this work we show that it is not sufficient just to 'consider' accessibility; in order to obtain meaningful results, the accessibility information must be used efficiently. For instance, although we find that the total free energy is a good criterion for selecting candidate binding sites, we show that it is not the best ranking criterion.

Despite many successful applications of existing target prediction algorithms (11–15,17,18,20,23–28), their specificity and precision remain far below 100%. This is especially due to the fact that the detailed molecular mechanism remains unknown. Initially, this lack of knowledge justified the development of empirical rules and scores such as those used in miRanda (11). Now that more complete knowledge has been accumulated (thanks to both experiments and successful earlier algorithms), it is possible to develop systematic approaches based on fewer empirical assumptions but invoking a careful statistical analysis that circumvents the ignorance of the detailed mechanism.

With the goal of increasing the precision of target predictions, especially when conservation information is not available, we introduce a new accessibility-based algorithm. This algorithm ranks predictions according to the over-representation of accessible complementary sites, and not according to the hybridization or total free energies. Whereas the free energy is only one of many contributing factors, the over-representation reflects directly the selection pressure on the coevolution of the 3′UTR and the miRNA, and therefore takes implicitly into account even the unknown factors affecting target selection. For two very different organisms (the fruit fly *Drosophila melanogaster* and the human) with large available data sets, the new method has a significantly higher precision than more elaborate standard methods while preserving the same sensitivity. Among the top 100 predictions in the fruit fly, our algorithm finds more than twice as many validated targets as do other accessibility-based target prediction methods (if their parameters are optimized), and at least five times as many if default parameters are used. In the human data set, our algorithm is the only one that outperforms the simple algorithm based on the seed 2–8 requirement. Interestingly, we found that the main reason for the success of the new method is not the folding algorithm used in our calculations; instead, the success is due to the efficient combination of secondary structure calculations to select accessible sites, and over-representation to rank the targets. In particular, our approach to include accessibility shows how to overcome the inherent limitations of algorithms that rank predictions according to free energy differences. Due to its features we called this method 'Prediction of Accessible MicroRNA Targets (PACMIT)'.

MATERIALS AND METHODS

Review of the statistical method

The full details of the molecular mechanism of miRNA target selection are still unknown. Instead of designing various empirical rules and scores, Robins and Press (31) and Murphy *et al.* (32) circumvent the ignorance of the detailed mechanism by a careful statistical analysis. The basic assumption is that biologically functional miRNA-target interactions arose by coevolution of the miRNA and its target. Therefore, complementary sites in real targets should correspond to over-represented oligomers. In addition, since functional miRNA-3′UTR pairs can arise either by a single strong binding site or by multiple weak binding sites (21,33), the whole of 3′UTR, rather than a single complementary site, is considered as a potential target. To compare the two possibilities fairly, this algorithm assesses the level of over-representation of one or more complementary sites, by computing a single hypothesis P-value (P_{SH}) for each miRNA-3′UTR pair. This gives an approximate probability that a given oligomer (or n-mer), complementary to the miRNA seed, is found by chance at least c times in the corresponding 3′UTR. The lower the P_{SH} is, the higher the chances that the 3′UTR is a functional target. Thus, for a particular miRNA-3′UTR pair, if l is the length of the 3′UTR and n the number of nucleotides in the seed, P_{SH} can be computed as

$$P_{SH} = \sum_{i=c}^{l-n+1} \binom{l-n+1}{i} P^i (1-P)^{l-n+1-i}, \quad (1)$$

where P is the probability to find the given n-mer by chance at any particular position in the 3′UTR. P is computed using a Markov Model (MM) based on the composition of the 3′UTR. The MM can be of order $k = 0, 1, \ldots, n-1$, depending on the amount of data available to compute the corresponding frequencies. (All our

predictions used $n = 7$, $k = 1$, and a separate MM model for each 3′UTR.) The general procedure is to compute P_{SH} for all possible miRNA-3′UTR pairs obtained from a set of miRNAs and a set of 3′UTRs, to produce a final list of predictions (i.e. miRNA-3′UTR pairs) ranked according to P_{SH}.

We included the accessibility of the 3′UTRs via the partial accessibility approach. Instead of looking for the number c of all complementary sites in the full 3′UTR of length l, we look for the number c_{access} of complementary sites in the 3′UTR that are also partially accessible. Additionally, instead of the full 3′UTR length l, the expression for P_{SH} must use the total number of partially accessible sites in the 3′UTR denoted by t_{access}. Altogether, the accessibility restriction is considered in the P_{SH} as

$$P_{SH} = \sum_{i=c_{access}}^{t_{access}} \binom{t_{access}}{i} P^i (1-P)^{t_{access}-i} \quad (2)$$

3′UTR and miRNA databases

The 3′UTR sequences for the fruit fly (release dm3 from 2006) were downloaded from the UCSC Table Browser (34) available at: http://genome.ucsc.edu. 3′UTRs for the human were obtained from the release 48 of Ensemble (35) available at: http://dec2007.archive.ensembl.org/index.html. After removing redundant 3′UTRs, we obtained databases of 10 803 and 19 447 unique 3′UTRs for the fruit fly and the human, respectively. The miRNA sequences (153 for the fruit fly and 885 for the human) were downloaded from miRbase v13.0 (http://microrna.sanger.ac.uk/sequences) (36).

Computing precision and sensitivity

In order to assess the precision (PR) and sensitivity (SE) of various methods, for the predictions in the fruit fly we used a set of 220 experimentally tested miRNA-3′UTR pairs, labeled as 'functional' or 'not functional' (Supplementary Table S1). This data set is based on the data set compiled by Kertesz *et al.* (24) and complemented with additional validated targets reported in miRecords (37). In the human we used the experimental results reported by Selbach *et al.* (4) in which protein levels were monitored under the expression of five different miRNAs. From these results, a data set of 15 806 miRNA-3′UTR was defined according to the fold change in protein expression: 2406 of these pairs were considered as functional and 13 400 as non-functional (8). We have used this data set to assess the precision and sensitivity of the predictions in the human.

Once the experimentally tested pairs are ranked according to the ranked list of predictions produced by a particular method, PR and SE of that method are computed according to the following formulas:

$$PR = \frac{TP}{TP+FP}, \quad (3)$$

$$SE = \frac{TP}{TP+FN}. \quad (4)$$

Here TP, FP and FN denote the numbers of true positives (functional pairs predicted as functional), false positives (non-functional pairs predicted as functional) and false negatives (functional pairs predicted as non-functional), respectively.

It is important to note that the data set used for the fruit fly is derived from the so-called 'direct validation' in which time-consuming experiments, usually motivated by computational predictions, are carried out to prove the functionality of each miRNA-3′UTR. Unfortunately, many negative results are not published. In contrast, the data set for the human (which was not inspired by computational predictions) comes from high-throughput measurements (known as 'indirect validation') in which several secondary effects might be involved in the up/down-regulation of proteins (8). All negative results are available.

Calculations with other methods

Our predictions were compared with the results of miRanda (11), PITA (24), IntaRNA (26) and RNAhybrid (14) with their default parameters and also with the following sets of parameters that were found to increase the precision. (i) miRanda: score cutoff ≥ 140, energy cutoff ≤ -20 kcal/mol, gap opening $= -9.0$ and gap extension $= -4.0$; target score (i.e. the overall score of a miRNA-3′UTRs pair in case that multiple binding sites occur) = sum of scores for all binding sites. (ii) PITA: we only considered sites with $\Delta\Delta G \leq -10$ kcal/mol, with seeds of length 7–8, not allowing mismatches or G:U wobble pairs; target score = PITA score for multiple binding sites. (iii) IntaRNA: we considered sites with $\Delta\Delta G \leq -10$ kcal/mol, with seed 2–8 allowing G:U base pairs; 3′UTRs were folded with RNAplfold (38) with a window size $W = 80$ and a maximum distance between paired bases $L = 40$ as recommended in ref. (19); target score = lowest total free energy. (iv) RNAhybrid: we considered sites with $\Delta G \leq -20$ kcal/mol, with seed 2–8 allowing G:U wobble pairs; target score = lowest hybridization energy. In all cases the ranking of the predicted targets was made according to the so-called 'target score' defined above. We emphasize that we did not limit ourselves to using only the default parameters preselected in each program, but instead tested various parameters. The parameters that had performed the best for a given algorithm were used in the second comparison with our method. Wobble G:U pairs were allowed in miRanda, IntaRNA and RNAhybrid because perfect seed matches are not supported by the programs. We strongly recommend using the optimized parameters listed above instead of the default parameters.

Ranking predictions according to the hybridization energy, total free energy, total accessibility and random ordering

In order to compare different ranking criteria, four different quantities were used to order the miRNA-3′UTR pairs containing at least one perfect match in the seed 2–8 region: (i) over-representation (PACMIT), (ii) the hybridization energy [with the program RNAplex (16)], (iii) the total free energy (with IntaRNA) and (iv) the total accessibility (with RNAplfold). By total accessibility we mean

the sum of P_{free}'s over all accessible 4-mers contained in all complementary sites. As for the hybridization and total free energies, if multiple binding sites are found in a given 3′UTR for the same miRNA, we considered the site with the lowest appropriate energy. For reference we also show the expected behavior for a random ordering of the seed-containing pairs.

Significance of predictions

A distinguishing feature of our algorithm is that besides a 'score' (P_{SH} in our case), determining a rank of a given prediction, it also estimates the statistical significance of the top N predictions when more than one miRNA-3′UTR pair is tested by computing the multiple hypothesis P-value P_{MH} (32). P_{MH} is computed from the number N_t of random genomes (with the same length and average dinucleotide composition) that produce lower P_{SH} values for their rank-t prediction than the real genome. Thus, for position t in the ranking,

$$P_{MH} = \frac{N_t}{N}, \quad (5)$$

where N is total of random genomes analyzed (typically 100).

An alternative measure of the reliability of the different methods is the false discovery rate (FDR), defined by

$$FDR = \frac{FP}{FP+TP}. \quad (6)$$

Since our algorithm provides single hypothesis P-values, FDR can be estimated following the procedure introduced by Benjamini and Hochberg (39). In particular, assuming independent or positively correlated hypotheses, the FDR up to the kth prediction can be estimated as

$$FDR(k) = \frac{m \cdot P_{SH}(k)}{k}, \quad (7)$$

where m is the total number of predictions and $P_{SH}(k)$ is the single hypothesis P-value for the kth prediction.

The statistical framework of the method also allows estimating the total number of targets, by computing the difference between the total actual and total Markov-model-predicted numbers of complementary accessible sites (where the totals are over all microRNA-3′UTR pairs) (31). E.g. for the 153 miRNAs tested in the fruit fly, our algorithm predicts that there are ~1900 targets with perfect complementarity to the seed at positions 2–8.

Thus, if one is interested only in 'strong predictions,' one should choose predictions with P_{MH} or FDR below a certain cutoff, e.g. 0.05, or only the top 1900 in the example of the fruit fly. However, for sake of comparison with other methods which predict a very large number of targets, we also considered 'weaker' predictions.

Supplementary Figure S1 shows the FDR computed from Equation (6) for several methods applied to the fruit fly genome. Also shown are FDR(k) for our algorithm (denoted Theor.) and the FDR [denoted Accessible seed 2–8 (random order)] obtained if the predictions of our method were randomly ordered instead of being ordered according to P_{SH}. By definition, the FDR for Accessible seed 2–8 (random order) is constant. We stress that unlike other methods, ours allows a theoretical estimate of the FDR, which may be useful in future applications.

RESULTS

Whereas over-representation is an excellent ranking criterion, hybridization energy is a poor one

Our algorithm ranks its predictions according to the over-representation of complementary binding sites among those which are partially accessible. In other words, assuming the coevolution of the miRNA and 3′UTR sequences, we expect that functional targets should contain complementary n-mers that are over-represented among the partially accessible n-mers (we use $n = 7$ for the seed length) (31,32). Over-representation is quantified by a single hypothesis P-value (P_{SH}) for each miRNA-3′UTR pair: the lower the P_{SH} is, the higher the chances that the 3′UTR is a functional target (see 'Materials and Methods' section for details).

Before including accessibility considerations into the model, we checked the performance of over-representation as a ranking criterion. For this purpose, we compared the precisions obtained when all miRNA-3′UTR pairs with at least one seed 2–8 perfect match were ranked by over-representation (our algorithm), total free energy, hybridization energy and random order (see 'Materials and Methods' section). Both in the fruit fly and in the human, over-representation performs much better than hybridization energy and random ordering. On the other hand, no significant difference was found between ranking by hybridization energy and randomly (Figure 1a and b). Total free energy yields higher precision than hybridization energy and random order, but performs only comparably (in the fruit fly) or much worse (in the human) than over-representation.

In addition, in Figure 1c and d, we show that ranking by over-representation places many more true positives among the top predictions compared to the rankings by total free energy, hybridization energy and random ordering. This is especially notable in the human, where among the top 5000 predictions, rankings by hybridization or total free energies find no true positives whatsoever, while ranking by over-representation finds 12 and random ordering finds approximately two true positives. This corroborates that over-representation is much more appropriate to rank the predictions than the more intuitive options like the hybridization or total free energies.

The observations shown in Figure 1a to d are the first major result of our work. Contrary to common belief, ranking predictions according to the hybridization energy does not gain much. Hybridization energy is only useful as a cutoff criterion, but not useful to sort predictions. The total free energy performs better, but still quite poorly for the very top predictions, at least in comparison with the over-representation. We believe that this is due to the complex evolutionary information captured by

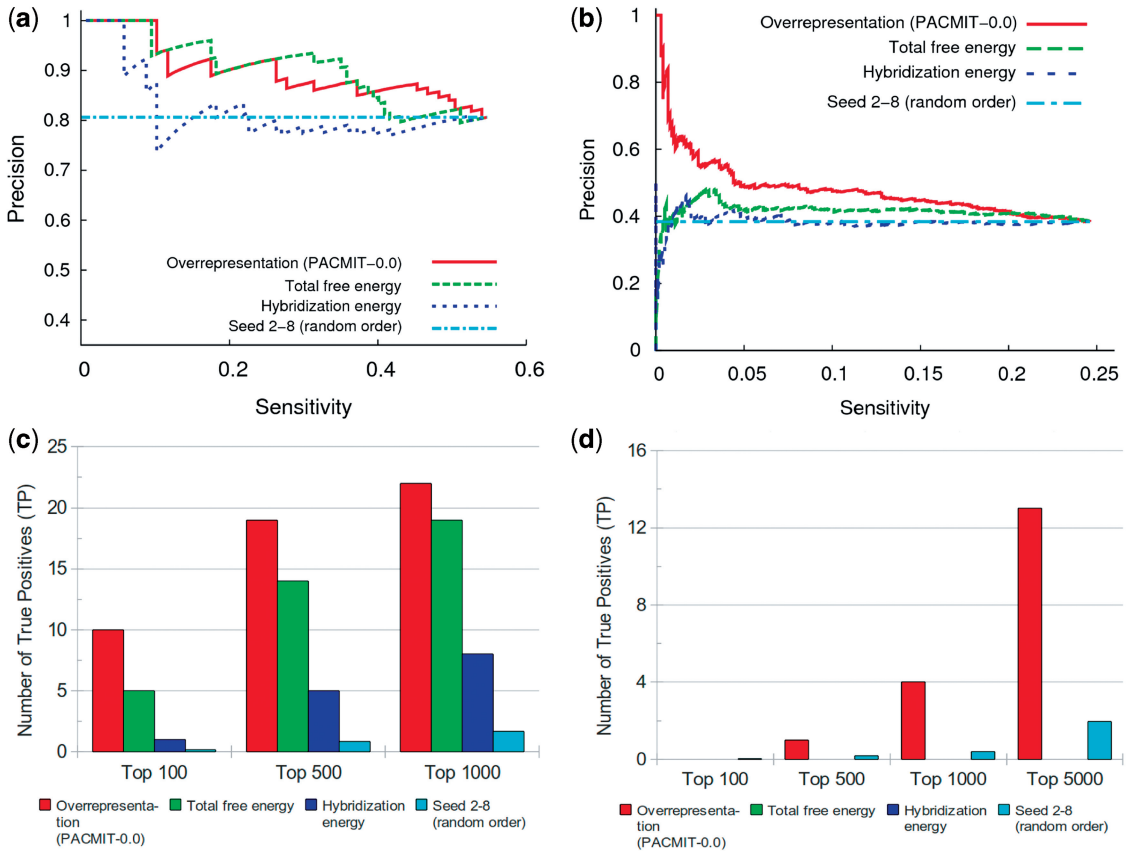

Figure 1. While over-representation is an excellent ranking criterion, hybridization energy is a poor one. All miRNA-3′UTR pairs with at least one perfect seed 2–8 match are ranked according to over-representation (measured by P_{SH} and labeled as PACMIT-0.0), total free energy, hybridization energy and random order. Precision versus sensitivity curves are shown for (**a**) the fruit fly and (**b**) the human. The number of true positives (i.e. experimentally validated targets) among the top predictions is also shown for (**c**) the fruit fly and (**d**) the human. In panel (d), the bars for hybridization and total free energies are not visible because the number of true positives for these two methods is always zero.

over-representation, but hidden from simpler physical criteria such as hybridization or total free energies. We note that the results of seed 2–8 (random order) in Figure 1 are already very good, and as shown below, much better than the results of algorithms based on hybridization or total free energy that omit the seed requirement.

An efficient way to include accessibility in the prediction algorithm

To simultaneously take into account the accessibility of the binding site and the possibility of multiple binding sites in the target 3′UTR, we included two modifications in the calculation of P_{SH}. Denoting by c_{access} the number of partially accessible complementary sites in the 3′UTR and by t_{access} the total number of partially accessible (but not necessarily complementary) sites in the 3′UTR, the P_{SH} value is given by Equation (2). As we will show below, this mechanism to include accessibility in miRNA target predictions is more efficient than currently used approaches. This is mainly because our algorithm incorporates the secondary structure to filter out false positives but not explicitly as a score: the ranking is given by over-representation (i.e. P_{SH}).

Selection of partially accessible sites

As almost no complementary site in the 3′UTR would be accessible at all times, less restrictive approaches based on partial accessibility have been proposed: only three consecutive nucleotides complementary to the seed (18) or four consecutive nucleotides complementary to any part of the miRNA (23) are required to be initially unpaired to allow the initiation of the binding process. Robins *et al.* (18) give two biophysical reasons for using a minimum of three free nucleotides: (i) three is the minimum number of unpaired nucleotides forming any RNA loop and (ii) base pair recognition between two strands needs at least three consecutive matches to form the double helix. We use a somewhat stronger condition, requiring at least four accessible nucleotides in the site complementary to the seed, but in such a way that the degree of accessibility required for such 4-mers is a flexible parameter.

In order to establish whether a given 4-mer within the 3′UTR is accessible or not, it is necessary to obtain the secondary structure by one of the available folding algorithms. The first attempts considered the minimum free energy structure (MFES) as a representation of the secondary structure (18,27). This approach has two problems: first, it neglects the possibility that the

microRNA binds to a 3′UTR structure with only slightly higher energy than the MFES. Second, the computed MFES is very sensitive to inaccuracies in the folding algorithms. Both problems can be avoided by considering the canonical ensemble of secondary structures (CESS) instead of the single MFES (23,24). Not only is using CESS more realistic, but it should also be robust against errors in the folding algorithms. The comparison of algorithms invoking either of these hypotheses with different algorithms based only on the hybridization energy and conservation filters suggests that both assumptions are reasonable. However, to our knowledge, the actual effect of either assumption has never been assessed explicitly on the basis of a common statistical framework. We make such an assessment here by comparing the results of our algorithm when the accessibility of the 4-mers is determined by either the MFES, or the CESS.

In case of the MFES, the 3′UTR is folded using the program RNAfold (from the Vienna RNA package) (40), and only unpaired 4-mers are considered to compute c_{access} and t_{access}. In case of the CESS, the question of accessibility becomes a matter of probability. Instead of asking whether a given 4-mer is free (i.e. fully unpaired) or not, one must ask for the probability P_{free} that this 4-mer is free. Therefore, we consider as 'partially accessible' any complementary site which contains at least one 4-mer with $P_{\text{free}} \geq P_{\text{cutoff}}$. The value of P_{cutoff} can be changed to obtain a given degree of specificity. We tried five different cutoffs: 0.1, 0.2, 0.3, 0.4 and 0.5.

Among the main concerns regarding the folding algorithms are worries that these programs cannot guarantee reliable structures for long sequences (41), and that proteins involved in the recognition of RNA strands are expected to hinder the formation of long-range interactions, thus making target accessibility a matter of local secondary structure (19). Therefore, in this work we used RNAplfold (also from the Vienna RNA package) to fold separately all possible subsequences of length W derived from a long sequence of length l, only allowing base pair interactions at a maximum distance L (38). For each 4-mer, P_{free} is computed as an average over all subsequences in which this 4-mer is present. This 'local folding' approach is much faster than folding the whole sequence at once ('global folding'), making it suitable for large 3′UTRs such as those of many human genes. In order to check the performance of the local folding against the 'naïve' global folding, we used P_{free} obtained from two different configurations: (i) global folding: $W = L = l$ and (ii) local folding: $W = 80$, $L = 40$ as suggested in ref. (19). The values of P_{free} are used to evaluate c_{access} and t_{access}, needed in Equation (2) for P_{SH}.

Canonical ensemble gives comparable or better results than the minimum free energy structure while local folding is not always better than global folding

To compare the effects of using the MFES or the CESS on the precision of target predictions, we compared PACMIT-MFES (PACMIT using only the MFES) with PACMIT-P_{cutoff} (PACMIT using the CESS with accessibility cutoff P_{cutoff}) based on both local and global folding. Since precision may be increased by simply sacrificing sensitivity, it is appropriate to compare the precision (PR_{max}) obtained with each method for common reference sensitivity, in this case the sensitivity obtained by PACMIT-MFES (Table 1). First of all, we observe (somewhat surprisingly) that for the fruit fly the global folding provides better results than the local folding (see column for fruit fly*). Second, we corroborate that the CESS usually (but not always) gives better results than the MFES. Third, the highest improvement in precision over PACMIT-MFES (if any), corresponds in most cases to $P_{\text{cutoff}} = 0.2$. For these reasons, in further comparisons we will use $P_{\text{cutoff}} = 0.2$ while we will employ global folding (labeled with an aterisk, as in PACMIT-0.2*) for the fruit fly and local folding (labeled without an asterisk as in PACMIT-0.2) for the human.

Table 1. Precision obtained with PACMIT using different folding schemes

Method	Precision[a]			
	Fruit fly*[,b]	Fruit fly[c]	Human*[,b]	Human[c]
PACMIT-0.0	0.900	0.900	0.483	0.483
PACMIT-0.1	0.923	0.900	0.469	0.486
PACMIT-0.2	**0.947**	**0.923**	0.466	**0.493**
PACMIT-0.3	0.947	0.900	0.414	0.483
PACMIT-0.4	0.923	0.878	N.A.	0.449
PACMIT-0.5	N.A.[d]	0.860	N.A.	0.424
PACMIT-MFES	0.923	0.923	0.405	0.405

[a]Precision obtained for the same sensitivity as that obtained by PACMIT-MFES, i.e. SE = 0.263 for the fruit fly and SE = 0.085 for the human.
[b]Using RNAplfold with global folding ($W = L = l$).
[c]Using RNAplfold with local folding ($W = 80$ and $L = 40$).
[d]Not available for the sensitivity cutoff.
*Used here and in the main text to denote global folding.

Accessibility improves precision

To find out whether the precision of the predictions is increased by including the partial accessibility requirement, we compared results obtained with PACMIT-0.2* (in the fruit fly) and PACMIT-0.2 (in the human) with the corresponding results obtained with PACMIT-0.0 (i.e. when all complementary sites were considered by setting $P_{\text{cutoff}} = 0.0$).

For all folding approaches the precision of PACMIT predictions in the fruit fly is always increased by the accessibility restriction (Figure 2a). Both CESS and MFES improve the precision of the method. As expected, we also observe that increasing P_{cutoff} (from 0.2 to 0.5) increases precision, albeit at the expense of sensitivity. In the human, however, considering accessibility does not lead to an improvement over PACMIT-0.0 (Figure 2b). While somewhat surprising, this attests to the strength of PACMIT-0.0 in ranking predictions rather than to the unimportance of accessibility. Altogether, we can conclude that accessibility considerations can significantly improve the precision of the method, and, in the worst case scenario, lead to similar results as with PACMIT-0.0.

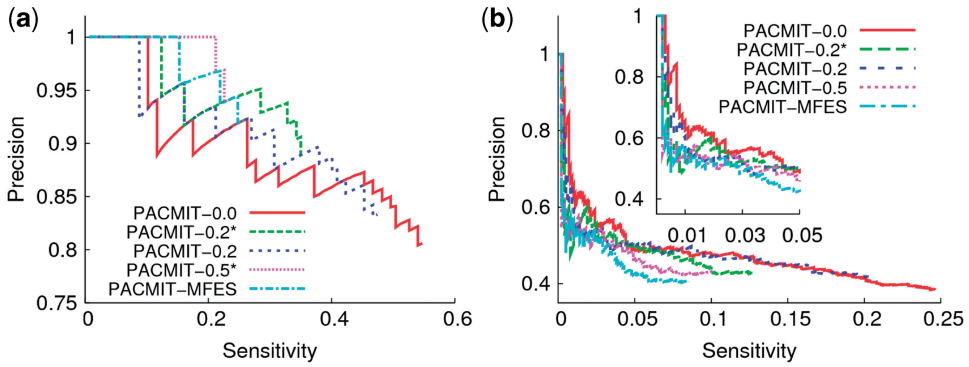

Figure 2. Considering accessibility can increase the precision of predictions. Precision versus sensitivity curves are shown for (**a**) the fruit fly and (**b**) the human. The different folding procedures used to include accessibility are compared with the case in which accessibility is not considered i.e. PACMIT-0.0. See the main text for the precise meaning of each label.

PACMIT shows higher performance than current methods using accessibility

We compared PACMIT with several currently available prediction methods. To the best of our knowledge, there exist three standalone tools using accessibility instead of conservation to increase the precision of miRNA target predictions: PITA (24), RNAup (25) and IntaRNA (26). From these three, we selected PITA and IntaRNA because IntaRNA and RNAup use a very similar approach and yield almost equivalent results while IntaRNA proved to be much faster than RNAup (26). In addition, we made comparisons with two methods that use neither accessibility nor conservation, but instead use the hybridization energy: (i) miRanda (11) because it is one of the methods most often used by experimentalists and (ii) RNAhybrid (14), which is only based on the hybridization energy and hence can serve as a reference to see the advantages of additional features included in the other methods. We emphasize that we did not limit ourselves to using only the default parameters preselected in each program, but instead tested various parameters. The parameters that had performed the best for a given algorithm were used in the second comparison with our method. We call these optimized parameters the 'high-precision parameters.'

In the comparisons we also include two reference curves: (i) as in Figure 1a and b, 'Seed 2–8' is the simplest miRNA target prediction algorithm in which all 3′UTRs with at least one perfect match to the seed are considered to be targets (predictions are randomly ordered). (ii) In the lines denoted 'Random' all possible miRNA-3′UTR pairs are randomly ordered. In other words, 'Random' corresponds to the absence of any algorithm whatsoever.

The precision versus sensitivity curves for the default parameters of the other methods are displayed in Figure 3a and b. In both organisms PACMIT shows the highest precision among all methods. In the human especially, the results of PACMIT are astonishing: its precision is about twice higher than the precision of the other methods (Figure 3b). Note that the other methods perform on par with the completely random ordering and much worse than the simplest seed 2–8 method (Figure 3b). In the fruit fly (Figure 3a) all algorithms except PACMIT and PITA have lower precision than the seed 2–8 method; those that neglect accessibility show precision lower than the completely random ordering. The results change for the high-precision set of parameters (Figure 3c and d) although PACMIT still has the highest precision. In the fruit fly, the other methods show a considerable improvement (Figure 3c). PITA has a higher precision than IntaRNA (presumably due to the requirement of perfect seed matches) and is the only method with a significant increase in precision also in the human genome (Figure 3d). As for miRanda, the apparent limitation of the thermodynamic model is well compensated by the heuristic score: not only is miRanda much more precise than RNAhybrid but it also shows a competitive behavior in comparison with PITA and IntaRNA. The poor precision of RNAhybrid in comparison with miRanda and with the random ordering confirms that efficient prediction of miRNA targets requires additional features beyond the hybridization energy. We note that by ignoring the seed requirement and ranking predictions based on the hybridization energy only, one can achieve higher sensitivity and also detect exceptional sites not satisfying the seed requirement. However, this is done only at the expense of a huge number of predictions and a very poor precision.

As the validation of the fruit fly data set has been historically motivated by the top predictions of existing computational methods that involve free energies in their ranking score (such as PITA), one would expect many true positives to be located among the top predictions of such methods. On the other hand, the validation was never motivated by a method ranking predictions by over-representation like PACMIT, and so one might expect fewer true positives among the top predictions of PACMIT. It is therefore astonishing that among the top 100 predictions, PACMIT-0.2* finds more than four times as many true positives than all other methods for their default parameters and more than twice as many for the high-precision parameters (Figure 4). Moreover, PACMIT shows the highest numbers of true positives

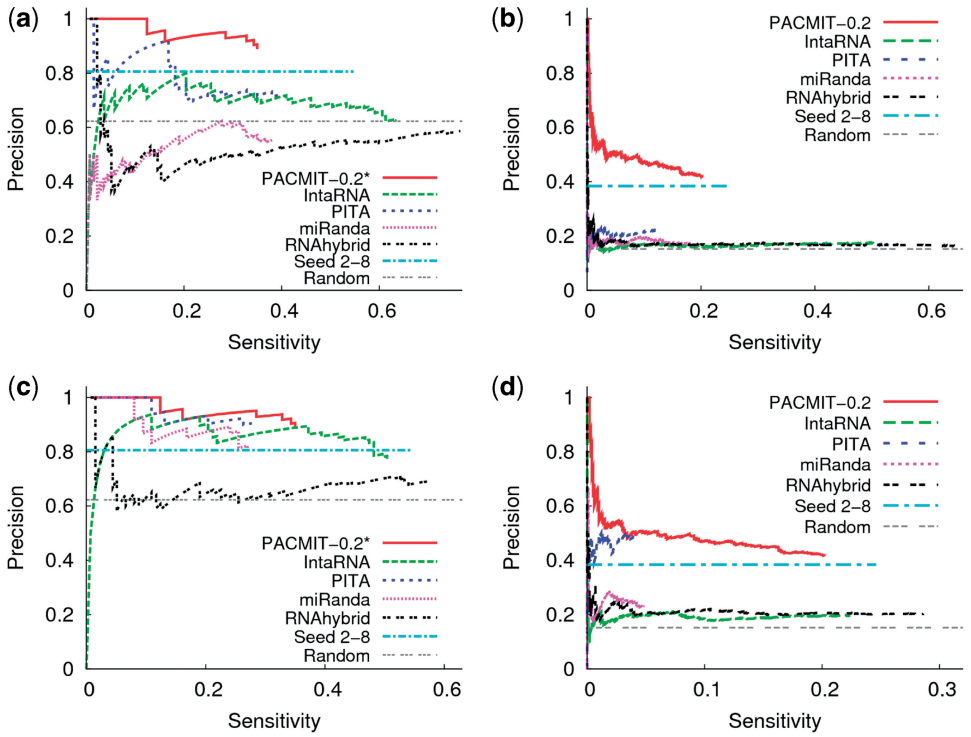

Figure 3. Comparison of PACMIT with other methods. Precision versus sensitivity curves obtained with the default parameters of different methods are shown for (**a**) the fruit fly and (**b**) the human. We also show the curves obtained with the 'high-precision' parameters for (**c**) the fruit fly and (**d**) the human. For the description of the 'Seed 2–8' and 'Random' curves, see the main text.

among the top 500 and top 1000 predictions in both situations (Figure 4). For the high precision parameters (Figure 4b), PITA shows better results than IntaRNA, but only comparable results to miRanda, which seems to capture in its empirical score more information than the thermodynamic calculations of both IntaRNA and RNAhybrid. Comparison of PACMIT and PITA in Figures 3 and 4 shows the advantages of over-representation as the ranking criterion. Although both approaches consider only perfect seed matches and rank the targets according to scores that account for multiple sites, over-representation enriches the top predictions with more true positives than the energy ranking of PITA.

Since experimental verification of microRNA targets is still not routine and since the numbers of validated or refuted predictions remain very small, experimentalists might be interested in comparing the number of validated targets that are predicted by a given method while precision remains perfect (PR = 1), i.e. before predicting the first target that has been experimentally rejected. In the fruit fly, the numbers of validated predictions while PR = 1 for default parameters were 21 for PACMIT-MFES, 17 for PACMIT-0.2*, 12 for PACMIT-0.2, 3 for RNAhybrid, 2 for PITA, and 0 for miRanda and IntaRNA (Figure 4a). For the high-precision parameters a great improvement is observed for PITA, especially due to a strict seed requirement (TP = 15), and for miRanda (TP = 11), but not for RNAhybrid (TP = 2) or IntaRNA (TP = 0) (Figure 4b).

Finally, Supplementary Table S2 compares the overall sensitivity and precision obtained when considering all predictions produced by each method. Note that this table completely ignores the ranking of predictions and hence, e.g. PACMIT-0.0 yields the same numbers as the simple Seed 2–8 method.

PACMIT is computationally more efficient than other methods

Besides higher precision, PACMIT also shows the lowest CPU time consumption. In the case of the fruit fly, using a single processor, we observed the following CPU times: IntaRNA (~50 h) >> PITA (~15 h) >> RNAhybrid (3 h) > miRanda (20 min) > PACMIT-0.2* = PACMIT-0.2 (10 min). PACMIT is faster than other methods because it does not require costly operations, such as the calculation of the hybridization energy, for each 3′UTR-miRNA pair. Besides the reusable 'fixed cost' of $O(\#3'\text{UTRs} \cdot l \cdot W^2)$ of folding 3′UTRs with RNAplfold, PACMIT has the computational cost of $O(\#\text{miRNAs} \cdot \#3'\text{UTRs} \cdot t_{\text{access}})$ because the only calculation that must be repeated for each 3′UTR-miRNA pair is the counting of perfect seed matches among the accessible sites, with the obvious cost of $O(t_{\text{access}})$. Another advantage of PACMIT is that all 3′UTRs in the data set are folded once and for all (without considering the microRNAs) and the information is saved in a database. One can then run the actual target prediction calculations varying P_{cutoff}, the length of the seed, and (most importantly) even the miRNA data set without having to rerun the costly

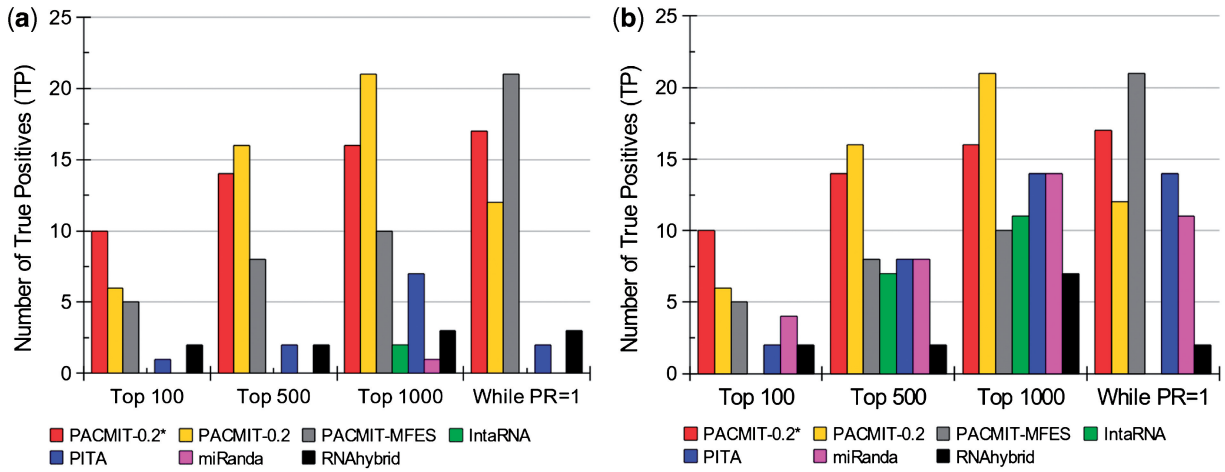

Figure 4. PACMIT has a higher number of validated targets among the top predictions. Numbers of validated targets among the top 100, 500 and 1000 predictions are shown for the fruit fly predictions by different methods under the (**a**) default and (**b**) 'high-precision' parameters. Also shown is the number of validated targets predicted before predicting the first false positive (see the rightmost cluster of bars, labeled 'While PR = 1').

folding algorithms. In other words, the amount of folding that is needed is the same for a single microRNA as for 100 or 1000 microRNAs. This feature is extremely useful also because the 3′UTR sequences are already known for most interesting organisms, while new microRNAs are constantly being discovered. Moreover, one might be interested in testing microRNA regulation across species boundaries.

As a specific example, folding all 3′UTR of the fruit fly took ~450 h (global folding) and 30 min (local folding), but the first calculation (for 153 microRNAs) with $P_{cutoff} = 0.1$ took only 10 min, the same time as that taken by the second calculation, now with $P_{cutoff} = 0.2$. This represents a major advantage over all methods in which thermodynamic calculations must be redone every time the algorithm runs. Due to this feature, we can say that PACMIT is a competitive option not only because of its precision but also because of its efficiency.

Effects of multiple-site scoring and of seed requirement on the precision of free energy based methods

We have also explored different ways to account, in a single target score, for the occurrence of multiple potential binding sites. While many free energy based methods ignore this issue altogether, a consensus is still lacking among the methods that do take multiple sites into account. Whereas PITA provides an empirical expression to achieve this, RNAhybrid suggests a formula but does not offer it in the publicly available program, and IntaRNA does not suggest any practical solution. To test the effect on precision of different schemes for multiple-site scoring, we used energies of all binding sites in a single 3′UTR to compute four different overall scores: (i) the lowest hybridization (or total free) energy, (ii) a PITA-like score (24), (iii) the arithmetic average of energies and (iv) the sum of energies. Surprisingly, we found that the PITA-like score reproduces almost exactly the results obtained with the lowest hybridization (or total free) energy. Additionally, despite the conceptual differences among the four tested expressions, the resulting differences in precision were almost negligible, except for the case of RNAhybrid in the fruit fly (Supplementary Figure S2).

Much larger effects on precision were induced by considering the seed. Requiring at least a partial seed complementarity (i.e. allowing G:U wobble pairs) considerably boosts the precision of a hybridization-energy based RNAhybrid and total-free-energy based IntaRNA (Figure 3). However, except for several top predictions, even the partial seed requirement does not lift the precision of RNAhybrid in the fruit fly above that of the random algorithm. Requiring perfect matches in the seed (instead of allowing G:U pairs) increases the precision of PITA by almost 100% in the case of the human (Supplementary Figure S3).

DISCUSSION

Using the concept of partial accessibility of the complementary sites in the 3′UTR and a successful statistical framework (31,32), we have proposed a highly precise microRNA target prediction method that does not require conservation information and is capable of considering multiple binding sites in a single 'score'. In our method, only complementary sites containing at least one accessible 4-mer are considered available for recognition by the miRNA seed. Two different approaches, the CESS and the MFES, were employed to determine the accessibility of 4-mers. We have observed that regardless of the approach used to include the secondary structure, considering accessibility increases the precision of the predictions.

Comparing values of precision achieved by various methods at multiple sensitivity cutoffs showed that: (i) using the CESS with intermediate P_{cutoff} values, instead of the MFES alone, further increases the precision of our algorithm (Figure 2 and Table 1), and that (ii) the

algorithm is significantly more precise than currently available miRNA prediction methods based on accessibility (Figure 3). Note that even PACMIT-MFES (based on the MFES only) outperforms PITA and IntaRNA, both of which use the CESS as well as more elaborate models to include accessibility (Figure 4). This comparison suggests that PACMIT uses the secondary-structure information more efficiently. We have observed that even when the precision of other algorithms is increased compared to the default settings by a strict seed requirement (as in PACMIT), the resulting precision for a given value of sensitivity is still lower than that of PACMIT. We have found that the ranking and therefore the quality of various methods is almost unaffected by the way in which multiple sites are considered, the lowest hybridization (or total free) energy being the most simple and efficient way to score the whole 3′UTR.

The thermodynamic approach used by methods like PITA, IntaRNA and RNAup, although reasonable, has important limitations: not considering the additional RNA–protein interactions taking place within the RISC complex directly affects the energy differences upon which target predictions rely. Hence, this negligence can not only change the ranking of predictions but also increase the rate of false positives. The PACMIT approach avoids such problems by not using the energy differences at all. Instead, accessibility information is used mainly to discard false positives and the ranking itself is done according to the over-representation of complementary sites among those that are accessible. Scoring of multiple sites is not treated empirically but at the core of the algorithm. To illustrate the advantages of different ranking criteria, we sorted all the miRNA-3′UTR pairs that have at least one partially accessible site (with $P_{cutoff} = 0.2$) according to: (i) over-representation P_{SH} (i.e. our algorithm PACMIT-0.2* for the fruit fly and PACMIT-0.2 for the human), (ii) total accessibility (see 'Materials and Methods' section for definition), (iii) total free energy, (iv) hybridization energy and (v) a random order (see 'Materials and Methods' section). The precision versus sensitivity plots in Supplementary Figure S4a and b show that in both genomes, ranking by over-representation gives the best results. As expected, ranking by the total free energy yields better results than ranking by the hybridization energy. Ranking by hybridization energy, in fact, performs comparably to a random ranking. In the fruit fly, only the total accessibility (which is a criterion also proposed in this work for the first time) performs on par with over-representation but in the human, over-representation is clearly the best ranking criterion. Moreover, in Supplementary Figure S4c and d we show that the number of true positives found among the top predictions of PACMIT is never outperformed by any of the other ranking criteria. Similarly, Figure 4 shows that PACMIT has the highest number of successful predictions before making the first mistake (i.e. the first false positive prediction).

These results show that limitations in modeling microRNA target selection may be partially circumvented by alternative statistical methods in which a detailed characterization of the miRNA-3′UTR interaction is avoided, i.e. by considering exclusively the 3′UTR length, composition, and accessibility. We have demonstrated that the free energy is not the most reliable criterion for ranking predictions, thus explaining why PACMIT-0.2* and PACMIT-0.2 showed better performance than PITA and IntaRNA. The success of PACMIT relies on the combination of two features: selecting candidates by accessibility and ranking them by over-representation. The achievement of PACMIT cannot be reproduced by considering accessibility alone, be it in the form of the total free energy or the total accessibility. Even if the total free energy is used to rank only the predictions with perfect complementarity to the seed, the precision of PACMIT remains higher. Similarly, miRanda performs better than RNAhybrid (and even than IntaRNA in some cases) because miRanda uses thermodynamics to select potential targets but relies on a different criterion to rank predictions.

Recognizing that conservation information (when available) provides the best filter to increase precision, our work in progress is focused on combining conservation and accessibility information. We are encouraged not only by the results shown in this work, but also by the performance of the 'bare' version of the present method, PACMIT-0.0, which ignores both conservation and accessibility information, yet, among the top predictions of the human data set performs comparably or even better than methods using conservation like PicTar, TargetScan 5.0, DIANA-microT v3.0 and EIMMo (Supplementary Figure S5).

The algorithm has been implemented in C. If interested in using PACMIT, please contact us.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

AKNOWLEDGEMENTS

The authors would like to thank Panagiotis Alexiou from the DIANA lab for providing the data set [used in Ref. (8)] of 15 406 miRNA-3′UTR pairs experimentally tested in the human, as well as the precision and sensitivity curves for PicTar, TargetScan 5.0, DIANA-microT v3.0 and EIMMo for this data set. We thank Didier Trono and Kathryn Maulaz for their useful comments on the article.

FUNDING

Funding for open access charge: École Polytechnique Fédérale de Lausanne.

Conflict of interest statement. None declared.

REFERENCES

1. Filipowicz,W., Bhattacharyya,S.N. and Sonenberg,N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, **9**, 102–114.

2. Ziegelbauer,J.M., Sullivan,C.S. and Ganem,D. (2009) Tandem array-based expression screens identify host mRNA targets of virus-encoded microRNAs. *Nat. Genet.*, **41**, 130–134.
3. German,M.A., Pillay,M., Jeong,D.-H., Hetawal,A., Luo,S., Janardhanan,P., Kannan,V., Rymarquis,L.A., Nobuta,K., German,R. *et al.* (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.*, **26**, 941–946.
4. Selbach,M., Schwanhausser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
5. Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.
6. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
7. Didiano,D. and Hobert,O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.*, **13**, 849–851.
8. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.
9. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
10. Mendes,N.D., Freitas,A.T. and Sagot,M.-F. (2009) Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res.*, **37**, 2419–2433.
11. Enright,A., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
12. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of Drosophila microRNA targets. *PLoS Biol.*, **1**, e60.
13. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
14. Rehmsmeier,M., Steffen,P., Höchsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
15. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
16. Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.
17. Maragkakis,M., Alexiou,P., Papadopoulos,G., Reczko,M., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K., Simossis,V. *et al.* (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.
18. Robins,H., Li,Y. and Padgett,R.W. (2005) Incorporating structure to predict microRNA targets. *Proc. Natl Acad. Sci. USA*, **102**, 4006–4009.
19. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
20. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.
21. Lai,E. (2004) Predicting and validating microRNA targets. *Genome Biol.*, **5**, 115.
22. Chan,C.S., Elemento,O. and Tavazoie,S. (2005) Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput. Biol.*, **1**, e69.
23. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.
24. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
25. Muckstein,U., Tafer,H., Hackermuller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
26. Busch,A., Richter,A.S. and Backofen,R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
27. Zhao,Y., Ransom,J.F., Li,A., Vedantham,V., von Drehle,M., Muth,A.N., Tsuchihashi,T., McManus,M.T., Schwartz,R.J. and Srivastava,D. (2007) Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell*, **129**, 303–317.
28. Hammell,M., Long,D., Zhang,L., Lee,A., Carmack,C.S., Han,M., Ding,Y. and Ambros,V. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods*, **5**, 813–819.
29. Obernosterer,G., Tafer,H. and Martinez,J. (2008) Target site effects in the RNA interference and microRNA pathways. *Biochem. Soc. Trans.*, **36**, 1216–1219.
30. Jinek,M. and Doudna,J.A. (2009) A three-dimensional view of the molecular machinery of RNA interference. *Nature*, **457**, 405–412.
31. Robins,H. and Press,W.H. (2005) Human microRNAs target a functionally distinct population of genes with AT-rich 3′UTRs. *Proc. Natl Acad. Sci. USA*, **102**, 15557–15562.
32. Murphy,E., Vaníček,J., Robins,H., Shenk,T. and Levine,A.J. (2008) Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: Implications for latency. *Proc. Natl Acad. Sci. USA*, **105**, 5453–5458.
33. Grimson,A., Farh,K.K.-H., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
34. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
35. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
36. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
37. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
38. Bernhart,S.H., Hofacker,I.L. and Stadler,P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
39. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.
40. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
41. Doshi,K., Cannone,J., Cobaugh,C. and Gutell,R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.